Karlsruhe Institute of Technology

Online Neural Network-based Language Identification

Master's Thesis

Daniel H. Draper

INTERACTIVE SYSTEMS LAB - INSTITUTE FOR ANTHROPOMATICS AND ROBOTICS

Contents

Motivation

Karlsruhe Institute of Technology

- ASR and SLT commercially great success ["OK, Google", Siri, Jibbigo]
- Best performing ASR models are still trained on one source language.[1]

- Performance increase for multilingual tasks[1], UI streamlining.

Motivation

- ASR and SLT commercially great success ["OK, Google", Siri, Jibbigo]
- Best performing ASR models are still trained on one source language.[1]

- Performance increase for multilingual tasks[1], UI streamlining.

Motivation

Karlsruhe Institute of Technology

- ASR and SLT commercially great success [”OK, Google“, Siri, Jibbigo]
- Best performing ASR models are still trained on one source language.[1]

- Performance increase for multilingual tasks[1], UI streamlining.

Specific Applications

Karlsruhe Institute of Technology

Lecture Translator

- Low latency, online application
- Small source language amount make performance increase likely

European Parliament

- Attempts to replace translators with SLT [2]
- Many languages challenging, great potential.

Specific Applications

Lecture Translator

- Low latency, online application
- Small source language amount make performance increase likely

European Parliament

- Attempts to replace translators with SLT [2]
- Many languages challenging, great potential.

Specific Applications

Lecture Translator

- Low latency, online application
- Small source language amount make performance increase likely

European Parliament

- Attempts to replace translators with SLT [2]
- Many languages challenging, great potential.

Specific Applications

Karlsruhe Institute of Technology

Lecture Translator

- Low latency, online application
- Small source language amount make performance increase likely

European Parliament

- Attempts to replace translators with SLT [2]
- Many languages challenging, great potential.

Related Work I

Historically

- single-language phone recognition followed by language-dependent, interpolated n-gram language modeling (PRLM)[3]
- multiple single-language phone recognizers and language-dependent parallel phone recognition (PPRM)[3]
- Gaussian Mixture Models (GMM)s [4]

Related Work I

Historically

- single-language phone recognition followed by language-dependent, interpolated n-gram language modeling (PRLM)[3]
- multiple single-language phone recognizers and language-dependent parallel phone recognition (PPRM)[3]
- Gaussian Mixture Models (GMM)s [4]

Related Work II

Modern Approaches

- Vector Space Modelling + PPRM[5]
- I-Vector approaches[6][7]

Related Work II

Karlsruhe Institute of Technology

Modern Approaches

- Vector Space Modelling + PPRM[5]
- I-Vector approaches[6][7]

Related Work III/Based on

Similar Approaches

- Matejka et. al[8] similar setup: BNF $\rightarrow$ LID, with averaging 5 LID nets.
- [9] M. Heck et al. evaluate LID approaches prospect of Lecture Translator integration: PPRM/PRLM + Hybrid.

- Markus [10]: "Language Adaptive DNNs for Improved Low Resource Speech Recognition", experimental setup very similar (LFV/LID)

Related Work III/Based on

Similar Approaches

- Matejka et. al[8] similar setup: BNF $\rightarrow$ LID, with averaging 5 LID nets.
- [9] M. Heck et al. evaluate LID approaches prospect of Lecture Translator integration: PPRM/PRLM + Hybrid.

- Markus [10]: "Language Adaptive DNNs for Improved Low Resource Speech Recognition", experimental setup very similar (LFV/LID)

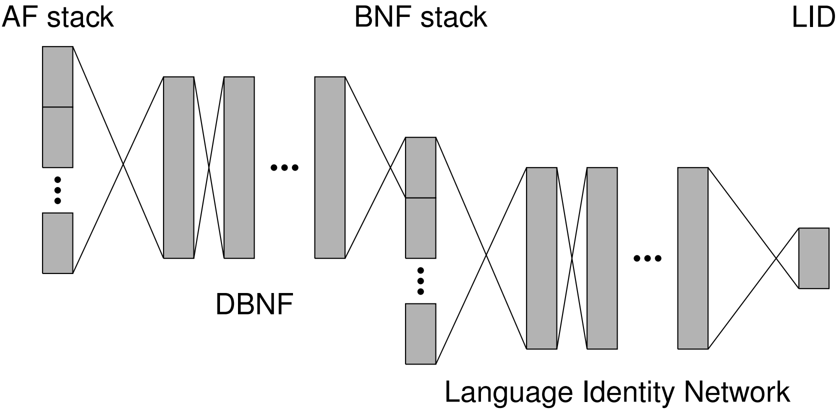
Related Work III/Based on

Similar Approaches

- Matejka et. al[8] similar setup: BNF $\rightarrow$ LID, with averaging 5 LID nets.
- [9] M. Heck et al. evaluate LID approaches prospect of Lecture Translator integration: PPRM/PRLM + Hybrid.

- Markus [10]: ”Language Adaptive DNNs for Improved Low Resource Speech Recognition“, experimental setup very similar (LFV/LID)

Experimental Setup



Data Sets

Karlsruhe Institute of Technology

Euronews 2014

- 10 Languages (EN, FR, DE, AR, ES, IT, PO, PT, RU, TR)
- Original Corpus: 72h / Language, Reduced corpus: 18h / Language (Random Sampling of 10.000 speakers)
- 80 % train data, 10 % each dev/test set

Lecture Data

- 3 Languages (EN, FR, DE) ≈10h per language.
- KIT lectures, InterACT25, DGA talks
- Validation of Euronews results, “right” patterns learned (different environment)

Data Sets

Karlsruhe Institute of Technology

Euronews 2014

- 10 Languages (EN, FR, DE, AR, ES, IT, PO, PT, RU, TR)
- Original Corpus: 72h / Language, Reduced corpus: 18h / Language (Random Sampling of 10.000 speakers)
- 80 % train data, 10 % each dev/test set

Lecture Data

- 3 Languages (EN, FR, DE) $\approx$10h per language.
- KIT lectures, InterACT25, DGA talks
- Validation of Euronews results, “right” patterns learned (different environment)

Data Sets II

Karlsruhe Institute of Technology

European Parliament

- 7 Languages (EN, FR, DE, ES, IT, PO, PT) 3.6h per language
- Simultaneous translation of all parliament speeches
- Further Validation, Proof-of-concept for Lecture Translator integration

Feature Extraction

Feature Definition

- Samplerate: 16 kHz
- Standard Janus Capabilities:
  - POWER
  - IMEL
  - PITCH
  - TON
- Context of 6 frames

Feature Extraction

Feature Definition

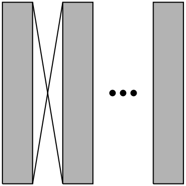
- Samplerate: 16 kHz
- Standard Janus Capabilities:
  - POWER
  - lMEL
  - PITCH
  - TON
- Context of 6 frames

Feature Extraction

Feature Definition

- Samplerate: 16 kHz
- Standard Janus Capabilities:
  - POWER
  - IMEL
  - PITCH
  - TON
- Context of 6 frames

DBNF

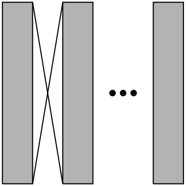


DBNF

DBNF

- AF Stack as input
- Targets context-dependent phoneme states
- 6 layers, each 1000 neurons
- BNF of 42 dimensions
- Stacked with context of 11 frames

DBNF

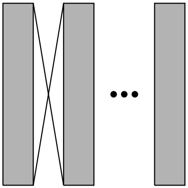


DBNF

DBNF

- AF Stack as input
- Targets context-dependent phoneme states
- 6 layers, each 1000 neurons
- BNF of 42 dimensions
- Stacked with context of 11 frames

DBNF

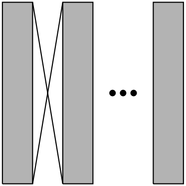


DBNF

DBNF

- AF Stack as input
- Targets context-dependent phoneme states
- 6 layers, each 1000 neurons
- BNF of 42 dimensions
- Stacked with context of 11 frames

DBNF

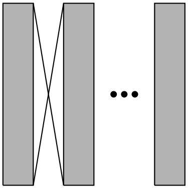


DBNF

DBNF

- AF Stack as input
- Targets context-dependent phoneme states
- 6 layers, each 1000 neurons
- BNF of 42 dimensions
- Stacked with context of 11 frames

DBNF

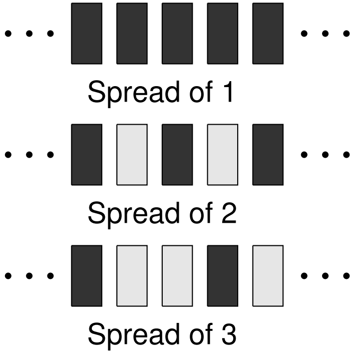


DBNF

DBNF

- AF Stack as input
- Targets context-dependent phoneme states
- 6 layers, each 1000 neurons
- BNF of 42 dimensions
- Stacked with context of 11 frames

Context Spread

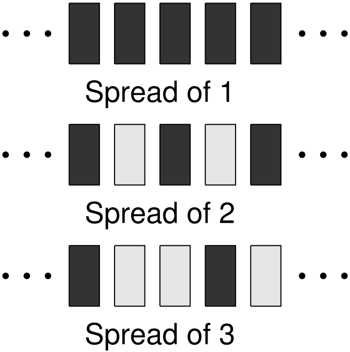


Spread Evaluation

- In [10], evaluated spreads of 2,3,6 for LFV
- Evaluation of spreads 4 and 5 for different nets on Euronews
- Spread of 3 outperforms 4 and 5, surprisingly 5 outperforms 4

Context Spread

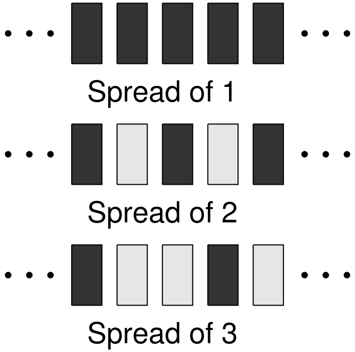
Karlsruhe Institute of Technology



Spread Evaluation

- In [10], evaluated spreads of 2,3,6 for LFV
- Evaluation of spreads 4 and 5 for different nets on Euronews
- Spread of 3 outperforms 4 and 5, surprisingly 5 outperforms 4

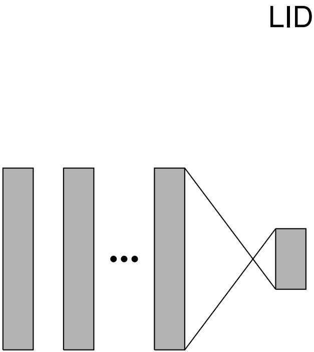
Context Spread



Spread Evaluation

- In [10], evaluated spreads of 2,3,6 for LFV
- Evaluation of spreads 4 and 5 for different nets on Euronews
- Spread of 3 outperforms 4 and 5, surprisingly 5 outperforms 4

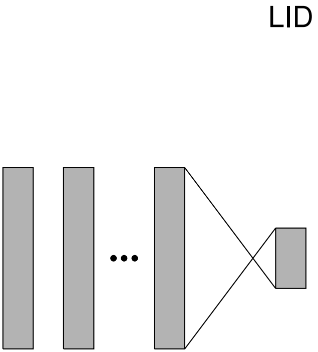
DNN Structure



Baseline Setup

- BNF with context as input
- 5 layers of each 1000 neurons
- Tanh activation, mse loss function
- Mini-batches size 2,000,000, learning rate(lr) 0.01
- Retrained with 1000:10 layer with lr 1

DNN Structure

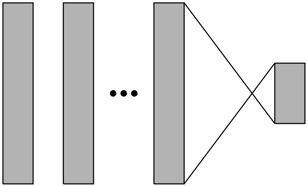


Baseline Setup

- BNF with context as input
- 5 layers of each 1000 neurons
- Tanh activation, mse loss function
- Mini-batches size 2,000,000, learning rate(lr) 0.01
- Retrained with 1000:10 layer with lr 1

DNN Structure

LID

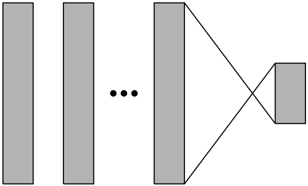


Baseline Setup

- BNF with context as input
- 5 layers of each 1000 neurons
- Tanh activation, mse loss function
- Mini-batches size 2,000,000, learning rate(lr) 0.01
- Retrained with 1000:10 layer with lr 1

DNN Structure

LID

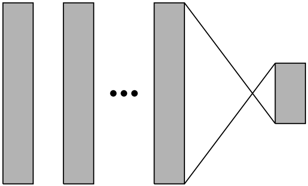


Baseline Setup

- BNF with context as input
- 5 layers of each 1000 neurons
- Tanh activation, mse loss function
- Mini-batches size 2,000,000, learning rate(lr) 0.01
- Retrained with 1000:10 layer with lr 1

DNN Structure

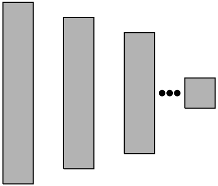
LID



Baseline Setup

- BNF with context as input
- 5 layers of each 1000 neurons
- Tanh activation, mse loss function
- Mini-batches size 2,000,000, learning rate(lr) 0.01
- Retrained with 1000:10 layer with lr 1

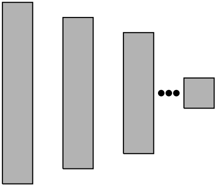
Evaluating Different Net Structures



Other Evaluations

- Lower LR ✓
- Baseline add Layers ✗
- Tree net with 5 layers ✓
- Tree net with 6 layers ✓
- Tree net with $>$ 6 layers ✗
- Cross-Data-Set training ✗
- Concat-Set training ✓

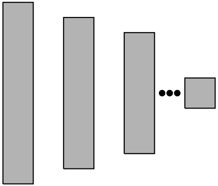
Evaluating Different Net Structures



Other Evaluations

- Lower LR ✓
- Baseline add Layers ✗
- Tree net with 5 layers ✓
- Tree net with 6 layers ✓
- Tree net with $>$ 6 layers ✗
- Cross-Data-Set training ✗
- Concat-Set training ✓

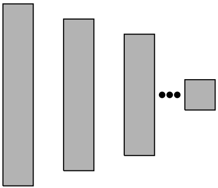
Evaluating Different Net Structures



Other Evaluations

- Lower LR ✓
- Baseline add Layers ✗
- Tree net with 5 layers ✓
- Tree net with 6 layers ✓
- Tree net with > 6 layers ✗
- Cross-Data-Set training ✗
- Concat-Set training ✓

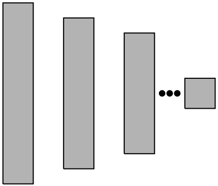
Evaluating Different Net Structures



Other Evaluations

- Lower LR ✓
- Baseline add Layers ✗
- Tree net with 5 layers ✓
- Tree net with 6 layers ✓
- Tree net with $>$ 6 layers ✗
- Cross-Data-Set training ✗
- Concat-Set training ✓

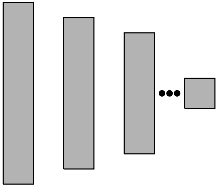
Evaluating Different Net Structures



Other Evaluations

- Lower LR ✓
- Baseline add Layers ✗
- Tree net with 5 layers ✓
- Tree net with 6 layers ✓
- Tree net with $>$ 6 layers ✗
- Cross-Data-Set training ✗
- Concat-Set training ✓

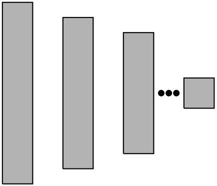
Evaluating Different Net Structures



Other Evaluations

- Lower LR ✓
- Baseline add Layers ✗
- Tree net with 5 layers ✓
- Tree net with 6 layers ✓
- Tree net with $>$ 6 layers ✗
- Cross-Data-Set training ✗
- Concat-Set training ✓

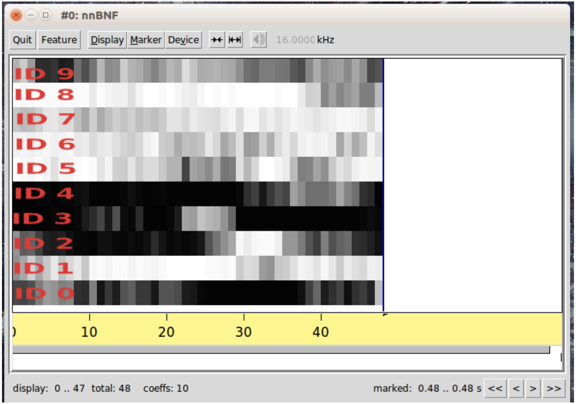
Evaluating Different Net Structures



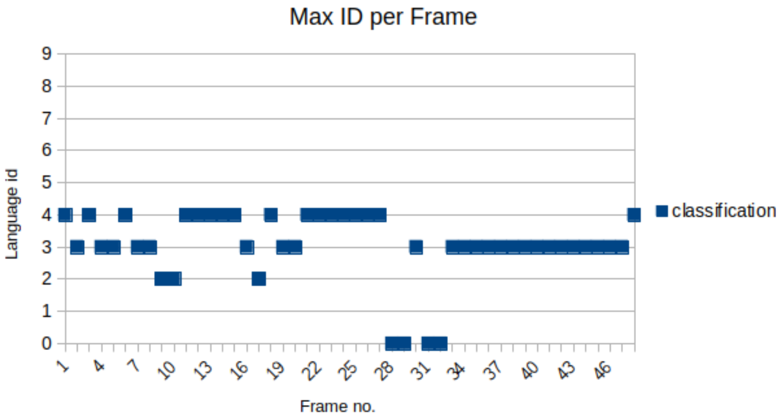
Other Evaluations

- Lower LR ✓
- Baseline add Layers ✗
- Tree net with 5 layers ✓
- Tree net with 6 layers ✓
- Tree net with $>$ 6 layers ✗
- Cross-Data-Set training ✗
- Concat-Set training ✓

Output


#0: nnBNF
Quit
Feature
Display
Marker
Device
16.0000 kHz
ID 9
ID 8
ID 7
ID 6
ID 5
ID 4
ID 3
ID 2
ID 1
ID 0
10
20
30
40
display: 0 .. 47 total: 48 coeffs: 10
marked: 0.48 .. 0.48 s

Output



Max ID per Frame
Language id
0
1
2
3
4
5
6
7
8
9
classification
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
46
Frame no.

Metrics

Error Rate

- Count outputs per language per frame
- Max count == actual language → correct, false otherwise
- Metric is percentage of error per language

Out-Of-Language-Error (OLE)

- Same counting as before
- Sum up FALSELY (as not equal to max language) classified languages
- Metric of “Sureness” of output
- Example: 50 frames classified as DE, 10 as EN, 5 as FR: $(5+10)/(5+10+50) = 0.23$
- Average OLE per language

Metrics

Error Rate

- Count outputs per language per frame
- Max count == actual language → correct, false otherwise
- Metric is percentage of error per language

Out-Of-Language-Error (OLE)

- Same counting as before
- Sum up FALSELY (as not equal to max language) classified languages
- Metric of "Sureness" of output
- Example: 50 frames classified as DE, 10 as EN, 5 as FR: $(5 + 10)/(5 + 10 + 50) = 0.23$
- Average OLE per language

Smoothing Filters

Approaches

- Counting Filter (count in window, in total as maximum only).
- Gaussian Filter (convolution with Gaussian kernel)
- Speech / Noise Filter
- Language Filter (If domain smaller than targets trained)
- Sequence Filter (longest sequence in window)
- Difference (Only count output if difference between 2 max is $>$ thresh
- Weighted Average (FILTER capability)

Results I

Filter	**Overall Error TEST**	**OLE TEST**
Bare Net	0.305	0.198
Gaussian Filter (WS 15)	**0.307**	0.176
Counting Filter (WS 100)	**0.307**	**0.006**
Best	**0.307**	**0.006**

Results II

- Euronews: Samples $>$ 500 ms: 16 % Error (10L)
- Lecture Data: 6.5 % Error (3L, concat with Euronews data)
- European Parliament: 35 % (7L, 3.5h/language)
- Post Processing: Counting, Gauss Filter promising, relative improvement OLE 97 %.

Demo (Easy)

Play

Demo (Easy)

Language	Number of Frames classified
English	1095
German	123
French	18
Total Error	0.11

Demo (Hard)

Play

Demo (Hard)

Language	Number of Frames classified
English	435
German	53
French	99
Total Error	0.25

Future Work

- RNNs [11]
- Post-Processing promising, further validation or actual focus
- Actual integration into online-environment like LT

Future Work

- RNNs [11]
- Post-Processing promising, further validation or actual focus
- Actual integration into online-environment like LT

Future Work

- RNNs [11]
- Post-Processing promising, further validation or actual focus
- Actual integration into online-environment like LT

Questions?

References I

- Z. Tang, L. Li, and D. Wang, "Multi-task recurrent model for true multilingual speech recognition," *arXiv preprint arXiv:1609.08337*, 2016.
- "Technology and corpora for speech to speech translation," http://tcstar.org/.
- M. A. Zissman *et al.*, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on speech and audio processing*, vol. 4, no. 1, p. 31, 1996.
- P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features." in *Interspeech*, 2002.

References II

- H. Li, B. Ma, and C. H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 271–284, Jan 2007.
- Y. Song, B. Jiang, Y. Bao, S. Wei, and L. R. Dai, "I-vector representation based on bottleneck features for language identification," *Electronics Letters*, vol. 49, no. 24, pp. 1569–1570, November 2013.
- L. F. D'haro Enríquez, O. Glembek, O. Plchot, P. Matějka, M. Soufifar, R. d. Córdoba Herralde, and J. Černockỳ, "Phonotactic language recognition using i-vectors and phoneme posteriogram counts," 2012.

References III

- P. Matejka, L. Zhang, T. Ng, H. S. Mallidi, O. Glembek, J. Ma, and B. Zhang, “Neural network bottleneck features for language identification,” *Proc. IEEE Odyssey*, pp. 299–304, 2014.
- M. Heck, S. StÃ$\frac{1}{4}$ker, and A. Waibel, “A hybrid phonotactic language identification system with an svm back-end for simultaneous lecture translation,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4857–4860.

References IV

M. MÃ¼ller, S. StÃ¼ker, and A. Waibel, “Language adaptive dnns for improved low resource speech recognition,” in *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, USA, September 8-12 2016.

J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno, “Automatic language identification using long short-term memory recurrent neural networks,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

References V

H. Li, B. Ma, and K. A. Lee, “Spoken language recognition: From fundamentals to practice,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, May 2013.

Theory of LID

- Given input $S = \{s(1), s(2), \ldots, s(T)\}$, $\{L_1, L_2, \ldots, L_N\}$

Formulation of LID Task[12]

$L^O = \arg\max_l p(S|L_l)$

With segmentation into phones υ (sequence Υ):

$L^O = \arg\max_l P(\Upsilon|L_l)$

With Viterbi decoding on set of phone models M:

$\Upsilon^O = \arg\max_\upsilon P(S|\upsilon, M)$

Combining last 2 equations:

$$\boxed{L^O = \arg\max_l \sum_{\forall\Upsilon} P(S|\Upsilon, M)P(\Upsilon|L_l)}$$

Theory of LID

Karlsruhe Institute of Technology

- Given input $S = \{s(1), s(2), \ldots, s(T)\}, \{L_1, L_2, \ldots, L_N\}$

Formulation of LID Task[12]

$L^O = \arg\max_l p(S|L_l)$

With segmentation into phones υ (sequence Υ):

$L^O = \arg\max_l P(\Upsilon|L_l)$

With Viterbi decoding on set of phone models M:

$\Upsilon^O = \arg\max_\upsilon P(S|\upsilon, M)$

Combining last 2 equations:

$$\boxed{L^O = \arg\max_l \sum_{\forall\Upsilon} P(S|\Upsilon, M)P(\Upsilon|L_l)}$$

Theory of LID

Karlsruhe Institute of Technology

- Given input $S = \{s(1), s(2), \ldots, s(T)\}$, $\{L_1, L_2, \ldots, L_N\}$

Formulation of LID Task[12]

$L^O = \arg\max_l p(S|L_l)$

With segmentation into phones υ (sequence Υ):

$L^O = \arg\max_l P(\Upsilon|L_l)$

With Viterbi decoding on set of phone models M:

$\Upsilon^O = \arg\max_\upsilon P(S|\upsilon, M)$

Combining last 2 equations:

$$\boxed{L^O = \arg\max_l \sum_{\forall\Upsilon} P(S|\Upsilon, M)P(\Upsilon|L_l)}$$

Theory of LID

- Given input $S = \{s(1), s(2), \ldots, s(T)\}, \{L_1, L_2, \ldots, L_N\}$

Formulation of LID Task[12]

$L^O = \arg\max_l p(S|L_l)$

With segmentation into phones υ (sequence Υ):

$L^O = \arg\max_l P(\Upsilon|L_l)$

With Viterbi decoding on set of phone models M:

$\Upsilon^O = \arg\max_\upsilon P(S|\upsilon, M)$

Combining last 2 equations:

$$\boxed{L^O = \arg\max_l \sum_{\forall\Upsilon} P(S|\Upsilon, M) P(\Upsilon|L_l)}$$

Theory of LID

- Given input $S = \{s(1), s(2), \ldots, s(T)\}$, $\{L_1, L_2, \ldots, L_N\}$

Formulation of LID Task[12]

$L^O = \arg\max_l p(S|L_l)$
With segmentation into phones υ (sequence Υ):
$L^O = \arg\max_l P(\Upsilon|L_l)$
With Viterbi decoding on set of phone models M:
$\Upsilon^O = \arg\max_\upsilon P(S|\upsilon, M)$
Combining last 2 equations:

$$\boxed{L^O = \arg\max_l \sum_{\forall \Upsilon} P(S|\Upsilon, M) P(\Upsilon|L_l)}$$